Soybean science blooms with supercomputers

16 August 2016, by Jorge Salazar

Ambitious SoyKB project aims to find and publicly share comprehensive soybean data achieved through the use of high-performance computing enabled by XSEDE. Credit: TACC

Knowledge of the soybean in the U.S. has come a long way since its humble start, namely as seeds smuggled by ship from China in the 1700s. A sanction back then from emperor Qianlong prevented trade outside of Canton. Undeterred, a former seaman with the East India Trading Company named Samuel Bowen first brought soybeans to Savannah, Georgia, in 1765. A couple of years later Bowen filed a patent for a new way of making sago (a starchy cake), vermicelli (noodles), and soy sauce from soybeans. Soybeans on colonial soil also got noticed by Benjamin Franklin, who wrote of their universal use in China as a cheese, which we now call tofu.

All the way through to the 20th century knowledge of soybeans came from the outside through selective breeding and manipulation of its environment—the warm weather, targeted water, loose soil, and full sunlight it needs to grow.

Today, an ambitious project called Soybean Knowledge Base (SoyKB) developed at the University of Missouri-Columbia (MU) aims to find and share comprehensive knowledge from within the soybean, its genetic and genomic data, all publicly available and achieved through the use of high-performance computing.

Dong Xu is one of the principal investigators of SoyKB, which he describes as a web resource for all soybean data from molecular data to field data including several analytical tools. Xu is a professor and department chair of computer science at MU.

"Our goal, first of all, is to provide a resource for people to find information about the soybean genes, their behavior, their gene expression, the metabolic pathways, and more," Xu said. He added that it's more than just a clearinghouse of data. SoyKB promotes deeper understanding through data analysis for scientists who want to improve crops to develop and verify their hypothesis. More than 2,000 unique users log on to the SoyKB website every month, and over 10,000 unique users have utilized SoyKB since it was developed in 2010.

SoyKB started small, initially focusing on the genomics aspects of soybean data, according to Co-PI Trupti Joshi. She is the director of Translational Bioinformatics at the School of Medicine Medical Research Office and assistant research professor in the Department of Molecular Microbiology and Immunology at MU.

"After a year or two," said Joshi, "we added the USDA germplasm data set, which gives you phenotypic information for about 19,000 soybean germplasm lines." Germplasm is basically the living

genetic information from seed banks scientists use to improve their breeding. "That is when we started building a lot of tools in the informatics suite," she said. These efforts, she added, are helping researchers find connections between the genomics data and variations in the germplasm lines.

"SoyKB has grown tremendously," Joshi said. "Over the years, we have had users from academic and industry environments. We have both domestic and international users from Canada, Brazil, India, China, and a lot of different countries in Europe. It's really been widely accessible." Times have changed since the days of American colonist Samuel Bowen.

The ultimate goal of SoyKB, said Joshi, is to improve soybean traits and support researchers in facilitating more enhanced soybean breeding techniques. "Our focus has been mainly on integrating multi-omics data sets about gene expression, protein expression, variations in the soybean, and then bridging it from this translational genomics side to the molecular breeding side, where it affects the soybean researchers and farmers," Joshi said.

The SoyKB project started its computation with NSF-sponsored XSEDE, the eXtreme Science and Engineering Discovery Environment, through an allocation awarded in 2014 on the Stampede supercomputer at the Texas Advanced Computing Center. In all, it has used about 370,000 core hours on a massive project to sequence and analyze the genomes of over 1,000 soybean germplasm lines.

The technique is called resequencing, where the genomic variations compared to a reference genome are found for each line. "The way resequencing is conducted is to chop the genome in many small pieces and see the many, many combinations of small pieces," said Xu. "The data are huge, millions of fragments mapped to a reference. That's actually a very time consuming process. Resequencing data analysis takes most of our computing time on XSEDE."

SoyKB sought the genetic markers for major soybean traits that include oil and protein content; soybean cyst nematode resistance; resistance to drought, heat and salinity; and healthy root system structure. "These data were very useful," said Xu, "because once we identified the genetic variations of those lines, they can be used for breeding purposes. It's really valuable data. In order to analyze the data, we didn't have enough resources. That's how XSEDE really helped us a great deal. In fact, we became one of the heavy users of XSEDE. Without XSEDE, we wouldn't be able to analyze this data. Now that the data are mostly analyzed, and we deposited this data into SoyKB, other researchers can also utilize it to answer questions of their interest," Xu said.

SoyKB was more or less a pipeline of Perl scripts when it first came to XSEDE, according to Mats Rynge. Rynge is a computer scientist with the Information Sciences Institute (ISI), part of the University of Southern California (USC). He's part of the XSEDE Extended Collaborative Support Services (ECSS) effort. ECSS is a pool of experts that help researchers use the cyberinfrastructure of XSEDE, a nationwide grid of some of the most powerful computational hardware and software in the world. Like the warm weather soybeans require, XSEDE provided the environment of hardware, software, and expertise SoyKB needed to thrive.

Rynge's group at ISI had experience with the Pegasus workflow, and he thought it would make a good fit for SoyKB to transform from scripts to a workflow optimized for supercomputers. One might think of Pegasus as the flow of water for a data-thirsty SoyKB platform. "Pegasus is a workflow system that can take a set of computational tasks, where one task produces a piece of data that is used by another task downstream," explained Rynge. Pegasus ensured that the ordering of the tasks was correct and that the data were formatted to best suit the execution environment of the parallel processing machines on XSEDE. It also handled the data management between tasks and the inputs and outputs.

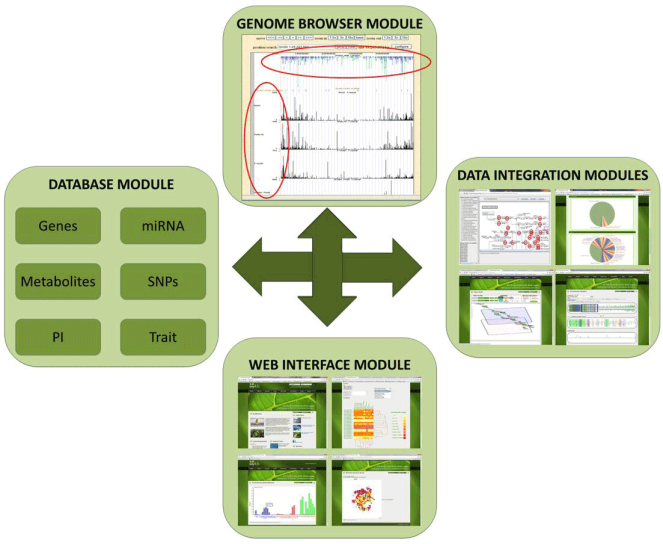
The workflow inputs were moved from MU and hosted on the data store of NSF-funded Cyverse. Cyverse, formerly iPlant, is a multi-institution resource for the life sciences to handle big data with platforms that provide data storage,

bioinformatics tools, image analyses, cloud services, APIs, and more. Cyverse resources supported the framework that allowed SoyKB to scale up for its thousand genome resequencing project. "For example, the data store framework really helped us tremendously," Trupti said. "We generated close to 25-30 terabytes of raw data from just one large-scale sequencing project."

Another move SoyKB took was to take its memory-guzzling genomic analysis from Stampede to Wrangler, a data intensive system that launched in 2015. Like the loose, fertile soil soybeans need, Wrangler's unprecedentedly large memory-to-core ratio gave ample room for the SoyKB workflow to avoid data bottlenecks. "I think part of the success story," said Rynge, "is when Wrangler came on, it turned out to be a much better fit. We transitioned from Stampede to Wrangler, and we have been very happy with it since."

"Many times our PGen Pegasus workflows would run anywhere from 10 to 15 days on the Stampede systems," Trupti said. "But then the same analysis could be completed in about 8 to 10 days when we moved those to the Wrangler system."

One big highlight of the SoyKB project is the easy-to-use suite of tools developed for informatics data analysis, said Joshi Trupti. "They are complete all the way from doing analysis with the soybean genome to getting you a view of what the gene expression might look like in different soybean tissues versus how certain soybean lines might respond to stress, whether it is in response to soybean cyst nematode worms or whether it is in response to drought stress. We actually built a system that stressed the user's perspective," Joshi said.



SoyKB is a web resource for all soybean data, from molecular data to field data including several analytical tools . Credit: SoyKB

MU scientists Trupti Joshi and Dong Xu were both on the team that in 2010 sequenced the first reference soybean genome. "It was exhilarating to be part of that community," Joshi said. "This was a great step forward for the soybean community with the first genome draft."

"Since then, we have actually had a second revision," said Joshi. "A version of the genome sequence and the gene model is being revised. We are really thrilled, because now we are in collaboration with Dr. Henry Nguyen at the University of Missouri and the Washington University genome sequencing center (McDonnell Genome Institute). We are sequencing the second reference genome for the "Lee" (PI 548656), which is representative of the southern cultivars. We are looking at a second reference genome coming out of soybeans," Joshi said.

Dong Xu of MU wants SoyKB to expand its platform to other systems through something like an 'app' store. "This means we have many individual tools other than the data analysis pipeline," Xu said. "We have a genotype-phenotype analysis pipeline. We also developed some visualization capacity. We have more than a dozen tools. We would like to

make these tools available to any other databases. We have been working with the corn community and others," said Xu.

Another future direction for SoyKB, Xu said, is to make it a genetic platform for other science groups to quickly develop their knowledge base. "Basically you could input the genome of any species and some annotations, and that would feed into what we call the 'KBCommons,'" Xu said. The KBCommons would generate websites automatically for scientists. "People can develop a knowledge base for a particular disease, like heart disease or diabetes," Xu said. "Even though there are a lot of databases for human genomics, there is still this need for these special purposes. Our platform can allow people to generate a specific platform quickly and easily."

One way that SoyKB is getting more users onboard is through an early research allocation on Jetstream, XSEDE's first scalable and fully-customizable cloud environment. The web-based user interface of Jetstream allows seamless integration with other XSEDE resources via Globus Auth.

With the help of XSEDE hardware, software, and expertise SoyKB has grown to be a rich ecosystem for the community of interdisciplinary researchers, industry, and nonscientists hoping to take advantage of the latest science on soybeans. And it has planted seeds of knowledge in the form of the many students that have participated in SoyKB.

"This is a great training environment for students," Trupti Joshi of MU said. "Being in an academic institution, where we have developed this system, it also gives a nice framework for us to be training the next generation of scientists. Plus, it gets high school students involved, even if they're simply interested in knowing what a soybean plant looks like and how it responds to stress. You could just go to the SoyKB website and do a quick search to look for one of the lines that are best for growing in a drought environment."

"One of the things that I really like about SoyKB when it comes to knowledge transfer is the student involvement," said Mats Rynge of XSEDE ECSS. "SoyKB had a more than normal number of students working with us. This is an important point, where the knowledge transfer is not just to computational scientists at some other project. It's really teaching students on how to do computing. That will hopefully help them with their computational needs in research when they are graduated and doing their own research."

Provided by University of Texas at Austin